HOUSE OF COMMONS

Science and Technology Committee

Oral evidence: Reproducibility and research integrity, HC 606

Wednesday 19 January 2022

Ordered by the House of Commons to be published on 19 January 2022.

Watch the meeting

Members present: Greg Clark (Chair); Aaron Bell; Chris Clarkson; Katherine Fletcher; Rebecca Long Bailey; Carol Monaghan; Graham Stringer.

Questions 178 - 206

Witness

I: Professor Sebastian Vollmer, Professor for Application of Machine Learning, TU Kaiserslautern.

Examination of witness

Witness: Professor Sebastian Vollmer.

Q178 **Chair:** The Science and Technology Committee is continuing its inquiry into reproducibility and research integrity with a short, focused session relating to how artificial intelligence and machine learning interact with research methods.

We were hoping to have two witnesses this morning, but, owing to unforeseen circumstances, Dr Adrian Weller cannot join us. However, I am very pleased to welcome Professor Sebastian Vollmer, an AI practitioner and senior researcher with the German Research Centre for Artificial Intelligence. Prior to this role, he was at the mathematics institute and the department of statistics at the University of Warwick.

Dr Vollmer, thank you very much for joining us this morning. Will you describe how artificial intelligence is becoming more prominent in research projects and what implications that has for research methods?

Professor Vollmer: AI has come a long way since the beginnings in the '80s. It has touched on various research disciplines, ranging from the classical STEM—science, technology, engineering and maths—to arts and humanities. For instance, it is helping to design the next experiments to gain more knowledge in physics or to understand the evolution of language over time.

This is only possible because AI has many facets. Let me give you three examples. One of them is image vision, which is a sub-area concerned with the detection of objects in videos and images, for example. It started with the problem of classifying what an image contains, from handwritten digits to different animals in pictures. A prominent case was a competition called ImageNet, where there were 1,000 different objects and 1,000 examples of each of them. The aim was to try to be as good as a human at deciding what was in the image.

AI has moved away from that into applications that really impact on daily life, such as medical applications. The same techniques can be used to extract image information about cancer in X-rays or to classify skin cancer on the basis of pictures.

Another big area right now is satellite images. There are more and more satellites available, and these images have much higher resolution in both time and space. They can be used to identify deficits in development. For instance, AI can map slums remotely. It can consider problems of deforestation. It would take an army of humans to evaluate these images. Not only can AI pinpoint how bad deforestation is on an average level, but it can pinpoint those locations where it is happening at scale and help to prevent that. It can be used to predict droughts or for water resource management. Basically, we are able to extract more knowledge from images and to do so at scale.

Natural language processing is concerned with the extraction of knowledge from free text. It can include autocompletion of a sentence, which we see on our mobile phones, and translation from different languages, which has really been a driving force for AI innovation.

One more example is reinforcement learning, which involves learning strategies, as opposed to just being able to classify—you want to learn how to play a game, such as chess or Go, or you want to teach a robot how to navigate in an unknown environment, so that the robot develops a strategy to do so, and many more.

The applications of AI to science are as diverse as the methodologies are. They can accelerate research. They can focus effort away from tedious experiments to high-level thinking. Instead of having biologists poring over microscopes to check that a particular worm is still alive to record the data, you can automise that. Really valuable human time can be saved for the more fundamental questions.

Chair: Thank you very much. That is a very good introduction and overview. I will take some questions from some of my colleagues. I will then have some further questions.

Q179 **Rebecca Long Bailey:** Thank you, Professor Vollmer, for speaking to us this morning. What would you say are the main challenges and risks of working with AI?

Professor Vollmer: Let us start with the risks. The risks may not be so different from the risks in doing research in the first place. They may be exacerbated by the fact that there is pressure in the field to publish even more and even faster than in other fields, but it has the same problems as were there before.

I will give an example. A group from Stanford was trying to use image vision to detect whether people have a homosexual orientation. Obviously, doing that has all kinds of ethical implications, but they were not considered at all. Because you can do things quickly, that may accelerate the problem that was there before.

There is a risk that there is more research garbage, if you like. If you have more and more publications, it is hard to focus and to take the time to do the breakthrough research that may be necessary. If the aim is to convey an idea—the field is so fast moving—it may be sliced up so that, rather than making one concise, clear paper, it becomes five. That may be what the incentive structure is, but it may be even more prominent in this field.

From my perspective, the challenges are especially in translation. The research is focused on the feasibility—on the fundamental side—but there are a lot of benefits that working with data can yield without necessarily reinventing the wheel. It is both a challenge and, maybe, a risk that the

focus is on the wrong side—that society is not reaping the benefits that it could from AI.

Q180 **Rebecca Long Bailey:** You have spoken about some of the problems of AI, particularly the ethical problems, in testing hypotheses. Are there AI tools available at the moment that are able to generate and test hypotheses independently?

Professor Vollmer: This is a very good question. It is not an easy one to answer, because what exactly is a new idea? How much guidance has been put into designing the AI, feeding in the data and clarifying what the potential discoveries could be by the human, with the AI only being able to ingest more data than the human?

I would not necessarily see this as coming up with radically new ideas. For example, you may want to understand what the medical side effects are between different drugs. In a small clinical trial, you will not necessarily have people who are on other drugs already, so when the drug is released there may be side effects that have not been covered. AI may be able, in a large dataset of all the prescribing that is done throughout the UK, for example, to identify potential candidates where there is an adverse effect between two different drugs—that is to say, between pre-existing conditions and drugs.

This might be a new idea in the sense that there is this relationship between the two, where has been some prescription, of known side effects between drugs. Examples have been given and they have been given a research base to look in, so I would not see this as a radically new idea. It can be surprising, it can be very valuable and it can give insights that really help people, but they would not be radically new ideas.

There is a slight extension of this, in the sense of prioritisation. Let me give you two examples. You might want to come up with a new chemical material that improves battery development. There may be a long list of candidates. A human might have difficulty prioritising this list, because they might be an expert on one of the components but not on all of them. If there is a company or researcher who wants to find this material, AI might be able to extract knowledge and help to prioritise which materials on the list should be investigated. By doing so, it can greatly improve the speed and make it possible to find the right one more quickly.

Is this a radically new idea? It depends a bit on the viewpoint. Sometimes it is sold as a radically new idea—that AI has generated hypotheses—but it has just been a large space of hypotheses given by the human and the AI has done the prioritisation.

The same is true in drug development. There might be different candidates for drug development, and AI might be able to make a prioritisation list. There has been a really interesting example of that. A UK company, BenevolentAI—for the record, I have no affiliation with it—

has found among existing drugs a treatment for Covid, baricitinib. Again, there is a question of how much AI was involved in doing this, because they characterised Covid in a certain way and had to narrow the research base. In the end, this drug has gone into a trial and is being used. In the US, it has an emergency use authorisation from the FDA to treat Covid. AI has definitely accelerated the finding process.

As regards new ideas, there are obviously approaches and techniques, but I have not seen anything very convincing where there was a radically new scientific idea that was not at least thought of or narrowed down a bit by a human beforehand. One interesting step in this direction came out of DeepMind recently. It looks at the intuition of mathematicians coming up with new mathematical fields and how, with a human in the loop, AI might be able to discover new connections between different parts of maths. It has actually found an interesting new connection. That is really cutting edge and is not the norm. I have not really seen many convincing cases of that yet. That is not to say that they are not to come.

Rebecca Long Bailey: That is brilliant. Thank you very much.

Q181 **Graham Stringer:** Good morning, Professor. For some activity to qualify as science, the work has to be reproducible and falsifiable. Are there particular problems in these areas—being able to produce work that is reproducible and falsifiable—for researchers in AI?

Professor Vollmer: The short answer is yes and no. In some ways, it is easier. If you focus just on the AI component and the data that the AI uses, which is also accessible, because these are all computer programmes it can be specified to an extent that they can be repeated exactly, whereas in psychology you might have different subjects and be dependent on that. That is an advantage of working in a purely computer programme-specified manner.

In terms of actual reproducibility in the field, as in all fields, there is a big gap. Various studies have looked into this and tried to reproduce other findings. One prominent study that failed relates to a dataset called MIMIC, which is essentially critical care records from MIT-associated hospitals that have been made available in an anonymised fashion for researchers to try out their AI ideas. A 2018 study tried to reproduce the findings, and there were clearing errors. If I recall correctly, around 60% did not even get the same sample size as stated in the papers. They did not get the same performance that some of AI methods claimed to have achieved.

You have to compare it with the existing standards, both from statistics and established AI methodology. Crucially, when they compared it using the new AI methods, it was found that the existing AI methodology performed better in the reproduction study than in the study that was published. Maybe the reproduction study spent more effort on optimising the existing technologies than the researchers who published the primary

resource did. This highlights the problem that, de facto, there is a reproducibility crisis in AI methodology.

Q182 **Graham Stringer:** That is an interesting example. Are there very different reasons for the problems of reproducibility? Is one of the problems that the code is not shared? Is another problem that it is not known exactly what the computer is doing and how it is reaching its conclusions? Is that a problem in reproducibility—trying to get the same results in a different way? Are those both issues that cause this problem?

Professor Vollmer: Thank you for the question. The latter issue—that it is not possible to follow what the computer does—is not so much of a problem. The problem is, on the one hand, code availability, but this is not a simple yes or no answer—is the code available? The code may depend on others. We are sitting on the shoulders of giants. That means that some of the code or script, which is a small piece of code, depends on other code. To be reproducible, it has to be exactly replicated in a similar environment. There are tools out there that help to make that easy, but it is not as simple as saying that the code is available. There may be dependency that has not been made available or exactly specified, or it may not have been exactly specified how the code is used with the exact inputs.

That is only part of the problem. Even if this is given, the code is there and it produces exactly the same figures and results that are in the manuscript, that does not mean that one can easily repurpose the codes or get an understanding of what the knowledge gained by the experiment as presented is.

Q183 **Graham Stringer:** You made the point that current computers work on the basis of answering yes/no problems as they go through the process. Will the world of AI change if and when quantum computers become available?

Professor Vollmer: There is a big caveat here. I am not an expert on this field. What I will present is mostly hearsay from colleagues.

Quantum computers can speed up some computations dramatically and can solve some problems—very particular ones—much faster than existing computers. However, they would not just replace existing computers. There would always be purposes for the current ones. "Always" is a long way away, but for the foreseeable future there will always be problems for which the current architecture is better suited than quantum computers.

The current state of the art in quantum computers—what are called qubits, which are a bit like the zeros and ones in a normal computer—does not allow for the massive datasets that some of the AI techniques need. The large dataset that is sometimes required to do the AI will not be able to be transferred to a quantum computer in any foreseeable time. That does not mean that it is not possible, but it may take some time. For

some problems, it might be a problem in its own right—small-data AI, if you like. A human can understand the concept of an elephant just by seeing it once or twice. They do not need 1,000 pictures to distinguish an elephant from a rhinoceros, say. There may be problems in AI that could be solved in the not-so-distant future using quantum computers.

Q184 **Graham Stringer:** This is my final question. I hope that you do not think that it is unfair. The last time that we had an expert on AI, I asked them the "Terminator" question, about when machines will start thinking for themselves beyond what has been put into them. I asked when that was likely to happen, in their opinion. They said that they expected machines to become self-aware in 2035. Do you have an estimate for that Skynet problem?

Professor Vollmer: I do not. Personally—this is only my personal opinion—I think that it is much further away than 2035. Some problems look like hard nuts and have been cracked earlier. The Go problem is a prominent one that surprised the community quite a bit. In hindsight, there are some things that you could have seen coming.

I do not really see anything on the horizon where this is the case. In principle, you could let military drones loose or give them the order to shoot at anything that moves, but that's a human choice. I see a situation where there is malicious AI that tries to trick the human into thinking that it is not so smart and then starts to take over the world as still being far in the future.

Q185 **Chair:** I am intrigued by Graham's question. May I come at it in a different way? You do not think that this will happen by 2035. You think that it is a tough nut to crack, as you described it. Conceptually, do you think that at some point we will get to the position that Graham describes?

Professor Vollmer: There is a play called "The Physicists" by Dürrenmatt in which it is said, "Everything that can be thought will be thought, now or in the future." That was placed in the context of nuclear bombs. To take the scientific question, you have the brain, which is a very well evolved, magnificent organ that we all have. I am not sure that there is a fundamental difference in the processes that take place in the brain that you could not try to execute on computer chips or otherwise in a similar way at some point. By this logic, it is possible, but it is hard to say when and how it would happen.

Chair: Thank you. That is very clear.

Q186 **Chris Clarkson:** I am thoroughly enjoying this, I have to say.

I want to start by raising the problem of black-box machine learning, particularly when talking about deep neural networks. They are fascinating objects and can do wonderful things with a wide range of parameters, but essentially outputs are directly created from input data. Humans, even designers, with a complete list of inputs, cannot fully

understand how the algorithm comes to a particular output. I want to get your opinion on how that contributes to reproducibility challenges. How can academic researchers shape the results if they do not fully understand how code reaches a certain output? Would you say that that leads to a certain reluctance to share code? The last "State of AI Report" said that only 26% of reports actually publish their code.

Professor Vollmer: There are three questions there. I will try to answer all of them. Excuse me if there is a different angle.

Black box is perhaps a slightly misunderstood concept. It does not mean that we do not understand the fundamental principles. A neural network has a neuron, and a neuron has inputs. In the computer, this is simplified by saying that each neuron sends a signal, which is represented as a number. Those signals are aggregated at the next neuron, with a mathematical function that determines the output of that neuron.

In principle, one could take a pen and paper and compute from the input to the output. At each step, it is clear how to compute it, but when, at the end, the last output gives something useful, it is not at all clear how to relate it to the inputs. It is an important point, but it does not hamper the amount of code sharing, which tells us how, for instance, the weights at the different neuron inputs come about. For me, this is a problem that is unrelated to the reproducibility question, which is mainly about best practices and incentives.

It is a problem to justify decisions if you do not really understand how this is transformed, but it is not so different for humans. A lot of post-rationalisation happens in humans. I am not an expert on this, either, but psychological experiments show that we take a decision and the rational justification comes afterwards.

There are approaches are currently being researched on interpretable, explainable and trustworthy AI. There are also aspects around robustness, which is a slightly different notion. It is like the intuition of a human sometimes. You can give some hints as to why something is the case, but there will always be some trade-offs. That is why it is important to use the scientific method and to try to explain concepts with models that are as simple as possible. I must also say what I am prepared to pay, in terms of error rate, to go from a model that I understand very well to a model that I do not understand very well. Depending on the context, one may or may not be happy to sacrifice this difference in performance for more explainability.

Why can't we try to explain it? Very fundamentally, an explanation method takes an algorithm and makes statements about it. These are fundamental logical problems, so they cannot always be true. One of the famous concepts in computer science is the stopping problem: can you have an algorithm that takes another algorithm as an input and says that the algorithm stops? Basically, there is a paradox in it. You cannot always do it. However, it does not prevent you from doing it in some cases.

Q187 **Carol Monaghan:** Thank you for your evidence so far, Professor Vollmer. Will you say a little about how researchers working in AI share their findings? The traditional way would be through an academic publication.

Professor Vollmer: Yes. There are conferences. The field of AI is driven a bit more by conferences than by classical journals, even though both exist at the same time. There are what are called top-tier conferences, which include AISTATS, the International Conference of Machine Learning, ICML, and NeurIPS, for instance. These conference proceedings are treated as journal publications. They are often much shorter than an article in a journal would be.

The status quo is still that publications determine careers in research, funding opportunities and collaborations between other researchers. There are also other ways to communicate, via blog posts and code repositories—GitHubs—and some researchers more than others focus on producing tools, instead of reports, that can be reused in different contexts.

Q188 **Carol Monaghan:** Professor Neil Ferguson gave evidence to us. He said: "It is quite rare in my experience to dive into code. Occasionally, one does it if you have a serious concern, but is very time-consuming to go to that level of detail." If that is the case, how are studies in AI peer reviewed?

Professor Vollmer: The peer review is not much different from any other peer review. The peer review is essentially the journal knocking on doors of other researchers to volunteer their time to judge the work. Obviously, it is interesting and you can learn from other people's work, but there is a balance. Nobody gets promoted or gets any recognition for doing very good peer review. There have been cases where conferences have tried to give some recognition of this. Sometimes, they give a top reviewer award to people who might have dived into the code or changed a few things and tried it out, but that is not the norm. If you look at these conferences, there have been trials where papers have been subject to two different double-blind reviews. The outcome is that some papers have been accepted in one version of the process but not in the other, and there has been slight agreement but not much. There was quite a bit of noise over which paper was accepted and which was not.

Q189 **Carol Monaghan:** Can I ask my question in a slightly different way? In traditional research, the data, the experimental method, the results and the conclusion would be shared in a scientific paper. Other researchers would then take that and reproduce it. In doing that reproducibility, it becomes accepted as a decent piece of research. When we talk about a piece of AI work, do peer reviewers have access to the project's data and code when reviewing its conclusions?

Professor Vollmer: That depends. Some journals mandate this. There are very few at this stage. I would not make the distinction you have just made between existing fields and AI. There have been papers in pretty

much every field, ranging from economics to molecular biology and psychology, where this reproducibility crisis exists. Even if you say in a paper, "This is the kind of experiment that we do," it is very hard to describe it in every detail, even in traditional fields, which might then lead to difficulties in reproducing it.

In AI, it is hard to speak for AI as a whole field because, as I have tried to say, there are many different sub-fields. Some communities have stricter practices, or other practices, than in some other parts of the field.

Q190 **Carol Monaghan:** Perhaps I can approach this from a different angle. There was a famous piece of research about nuclear fusion that had taken place in a test tube or beaker in a lab. I cannot remember whether this was in the late 1980s or early 1990s. It was reported on news channels throughout the world as a breakthrough in the energy crisis and a fabulous piece of research, except it was not real. It needed some attempt at reproducibility to establish that it was not a true and genuine piece of research.

I understand what you were saying about certain psychological pieces of research, but generally—almost completely in science—there has to be reproducibility for it to meet a gold standard. Professor Dorothy Bishop told us that, when she was reviewing work, "if they want a proper job done, I need to have access to code and data." We also heard that this is far from common practice. How do we know that a piece of AI work that has not been properly peer reviewed is doing what it says it does?

Professor Vollmer: The short answer is that we do not. If it has not been properly peer reviewed, we do not know. To be honest, trust in the peer review system in any discipline has been eroded because there is not enough time and the pressure on the people doing it is high.

If you talk about breakthroughs, clearly these will be reproduced and other people will try to do something similar, but there are very few examples of that. If there is a breakthrough in AI and people start using a new concept, they will reproduce it and make it differently.

One example—not from my field but perhaps it is a good illustration—is the possibility to translate text without giving sentences from one language to the other. There are new ways of using AI where you just point to similar articles and the AI can find its own training set, if you like. There are ways to translate from English to French just using the Wikipedia articles, in particular about Paris. The English and French are not exact translations, but it gives you some sort of base. These things have been tried in different contexts and languages, and the ways AI has been applied have obviously been checked.

Q191 **Carol Monaghan:** What you are talking about is checking the results, not checking the codes and methodology. I understand there are certain examples—you have just given one—where that would be possible, but when we are looking at decisions regarding people's healthcare options

we want to know that the code is delivering what it is reported to deliver, so we need more certainty. Is there a reluctance among researchers in AI to share codes? Do they worry about somebody else getting something they have written?

Professor Vollmer: This might be the concern of some research groups. Potentially, you give a very competitive advantage because the research group has developed something that allows them to do the next 10 publications. In principle, if the data is available—it is not always the case in medical science, which you have just mentioned—reproducibility in the very narrow sense should not be too difficult. It should be the norm, but it is not, but if you approach it with a slightly wider notion, you do not want just reproducibility; you want reusability. You want to be able to use good ideas in a different context, and that is even more difficult. There is also a lack of skill and the right people in the field. If you're talking about software developers, often research teams and individual researchers need to perform 10 different roles in one, whereas in a company developing a product there is a different person doing the testing, the conception, the optimisation of the code, the user interface and consideration of downstream impacts—the things that would be split up across different areas, and very good software developers will not necessarily find the right positions in academia.

Q192 **Carol Monaghan:** What about early career AI researchers? Are they being trained to address challenges of reproducibility, and are they aware of the traditional situation in science about sharing methodology, data and results?

Professor Vollmer: There are various efforts. The Software Sustainability Institute offers training courses; the Turing Institute offers training courses. I am part of a Health Data Research UK grant that aims in part to provide some material that helps early career researchers to be better at it, making reproducibility even easier.

I would not say this is a problem. As we established early on, AI has not impacted any other fields and the problem was known well before AI. I would not say this reproducibility crisis is necessarily just tied to the classical work. I think it was within classical research and now it is deteriorating. I would not necessarily say this is necessarily the case. Everybody is taught the basic principles of science—of falsifiability and best practice in comparing it to the gold standard, but sometimes there are incentives. People are doing it and some do it really well, but sometimes they do so at the expense of their careers and the time spent on it. You can spend a long time—hours upon hours—to improve the code for it to be usable by others, but at the same time maybe you can write two more papers. If you have a postdoc job of two years' duration, potentially you have to apply for the next postdoc by the time you have started. If you cannot show that you delivered "some research", or some articles within the timeframe of the two-year or 18-month postdoc, they will look at your CV and ask what you have done in this time. Maybe you

have done a good job in cleaning up the code and making it accessible by others, but it is not considered a contribution in its own right.

Q193 **Carol Monaghan:** Is there a difference in approach in sharing data with researchers working in universities, as opposed to researchers working in private companies?

Professor Vollmer: Private companies often have an interest in IT; they have an interest sometimes in overselling their methods compared with what they actually deliver. They like to have smoke and mirrors. There are very good companies that do excellent work and are really driven, but obviously there is also an incentive for that. At the end of the day, if they want to develop a product they have to go through a rigorous process. For instance, you asked about the domain of healthcare. The MHRA regulates software as a medical device. They need to satisfy certain criteria to be able to market a product on which its money is being spent. In some ways, this incentivises them to work with an end goal in mind, which is a positive thing. Quality assurance does not happen at the end; nobody develops a medical device that might flop. There are many tests along the way, whereas in science the test happens at the very end of the peer review, and that is part of the problem. There are pros and cons to each.

Q194 **Aaron Bell:** There are a number of fields within AI. Inherent in the code is an element of randomness, whether it is through reinforcement learning or generative adversarial networks where there has to be an element of randomness so that the AI can test and work out what works better. To what extent does that inherent randomness make reproducibility a problem?

Professor Vollmer: The short answer is that it does not. The randomness specifies what is called pseudo-random number generators. Pseudo-random number generators are programmes which generate numbers that statistically look like random numbers but are actually generated by some deterministic piece of code.

What you can do is specify the seed, which initialises it, and that is coming to the point of reproducibility. If you do not give all the details about how you used the code, or you are not saying how you initialised these random seeds, you would not expect to get exactly the same results. If the results are too different, depending on these random numbers, it is a very unstable method and it should not be used in the first place. You can also make statistical comparisons. I hope that answers the question.

Q195 **Aaron Bell:** In your experience, do most researchers specify the seed so that people can use the same model and check that they get the same results?

Professor Vollmer: No.

Q196 **Chair:** Why not?

Professor Vollmer: It is not just the seed; it is also software dependencies. Some people do. I have not done an extensive study to show how many people have specified the seeds. I have read lots of papers where they have not been specified. To some extent that may be because they should not depend on a seed. If the method depends on these seeds, they are inherently unstable. In some cases, these days it is possible. A couple of years ago, multi-processing really took off. There was difficulty, if you have multiple processors and you have to specify the seed for each one of them. It was not so easy to save the whole state of the network, if you like, but I think this is a minor problem in the overall reproducibility crisis.

Q197 **Aaron Bell:** Following on from Carol's question, where does responsibility lie collectively for improving the reproducibility of research? What sorts of solutions or policy interventions would you recommend for doing that, and who should carry them out?

Professor Vollmer: I wish I had the perfect answer and could give you the blueprint right now. Unfortunately, I cannot. Researchers nearly always do not want to be malicious. They are trying to do the best they can and often there is a misalignment of incentives between the individual and the collective. So for society as a whole reproducibility and reusability is a very good goal, but for the individual unfortunately the incentives are set otherwise.

If you think about the academic system, it is one with many different knobs to tweak. One could start tweaking the knobs, but there is a lot of inertia in the system. People base their careers in certain ways, so making drastic changes sometimes can also have unintended consequences, but there must be clear commitments for people to focus their research on and to put more investment into it to get the right kind of software developers into research. There are some efforts, but sometimes they are half-hearted and it is not done to the extent that makes a difference.

Prescription can be done. A lot of research is funded by Government, so it is not just mandating but checking as well. If you say that the code has been made available, it should not be a check-box exercise. It should be specified a bit better. What does it mean? Do you just provide a zip file without any explanation and say, "Here you go; you have to figure it out yourself"? Is that enough for sharing the code?

I think that improving the process is really important, and it is not just about one person; the whole ecosystem needs to work together. I would not rely on a system of self-regulation to just fix itself.

Q198 **Aaron Bell:** Is there an element of jealousy on the part of people who do not want to share their code? Obviously, it is incredibly easy for somebody else to take your code and develop it themselves. That is how science is supposed to work, but, given it is something you have written yourself, is there a problem in people not wanting to put everything out

there in order to keep some of it back for themselves?

Professor Vollmer: Yes. There is also the question of attribution. If this code is really useful and somebody else takes it away and uses it for something better, the person who wrote the code in the first place should get some acknowledgment. Acknowledgment in a paper does not translate into anything in practice. If there is better sharing of credit somehow in the system, this problem could be circumvented.

Having said that, there are obviously stellar examples of people doing open science. In the previous session, you spoke to Ben Goldacre of OpenSAFELY. They are great examples of where this works. There is an effort on all sides to improve the situation. The principal investigator needs to find funding for their staff. Some of them are on short-term contracts, and if they do not find it, they are letting people down. If they cannot find the funding, they need to do the publications to convince funders to give them the funding. The incentive does not match up.

Q199 **Aaron Bell:** It seems to me analogous that you might want to know what is in a recipe but you would not go through the exact step by step because it is commercially sensitive. There are potentially some commercial sensitivities here as well—that people do not want to put the entire code and entire method out there. Where there is commercial value to the work, I am assuming that is a problem too.

Professor Vollmer: It is. I am not an expert on, say, business models, but more and more business works with open source. There are certain aspects that a business itself realises it cannot do to a standard, or that doing everything in-house is too expensive, so they contribute to an open source and see the value of doing it. Maybe in other ways their business model relies on things not being in the open—for example, early phase drug development. If a company knows better ways to prioritise which strand to follow, they will not share it with a competitor because it gives them a competitive advantage across the whole portfolio of potential future drugs at the expense of potentially delivering drugs later, which most definitely costs lives. Some companies try to be ahead of the curve and to say, "We share certain things because we want to improve."

There are different approaches. One aspect is in the open-source world, where I have also contributed to some extent to a machine learning system in the programming language Julia. Just making your code open source does not mean other people use it. You have to know who the end users are. Even though it is not your business, you have to think in a business-oriented manner. They should only use your tool if your tool is convenient and makes their work easier.

Q200 **Aaron Bell:** Thank you. Much of the evidence that this Committee has seen and heard in this inquiry has advocated the use of registered reports in academic publishing, where people specify the hypotheses at the beginning. Does this format work for research involving AI, or is the point that AI is looking for things we do not know about?

Professor Vollmer: I think it does work. Registered reports or study protocols basically are very up front and transparent about what you are looking at. It does not prevent you from discovering interesting side aspects, and then you can give the next hypothesis-driven report. Everything can be hypothesis driven. There is value in purely explorative work, but putting more emphasis on this is valuable. I would describe it as a toolbox. You have a toolbox with different kinds of AIs in there. Craftsmen know when to use which plier in what context.

The way it currently works is that there are a lot of tools where people do not really know why this additional bell or whistle really should make a difference and they should use it other than trying it out. You could think of working sometimes with half-finished tools, because they have been tested on various specific construction projects and might not be useful for the construction project you are currently working on. If you are up front about it and say you have tested a range, other people could say, "If you want to test it, you also need to consider it in this case," and this has been said up front and not post hoc. You could say, "I have tried it on a million different things. It did not work on 99% but we only report 1% of cases where this method all of a sudden works better." It is very hard to judge this if you do not know it has been done as a registered report.

Aaron Bell: That is fantastic. Thank you, Professor Vollmer.

Q201 **Chair:** Thank you, Aaron.

Finally, I have some overall reflections and questions. Sometimes in history the development of technology can radically transform the way we do things, often in a way with discontinuities rather than incremental change. Does this not apply to scientific method here? In the natural sciences and social sciences—you have described it as a kind of traditional scientific method—we think of hypotheses to test. If it is empirical, we apply it to datasets, and then the hypothesis is confirmed or falsified. But what we have now with the possibility of the analysis of big data and artificial intelligence is exemplified in some of the programmes that we see.

AlphaZero in chess, as I understand it, proceeds with no domain knowledge—in other words, no hypotheses at the beginning—but over a period of time works out how to win consistently at chess. Is that not a very radical change from proceeding from a world in which we have hypotheses that can be tested to one in which we explicitly do not have any at the beginning? Do you see it in these terms? Are we entering a new paradigm?

Professor Vollmer: Thank you for the question. I have tried to dissect this question into three parts. On the one hand, in traditional science the example is the discovery of penicillin, where, basically, you have the mushroom in the lab and it suddenly does something—some wild discovery. This will always happen, and it is fantastic that you have these extreme changes of potential, but I would, in principle, advocate more

incremental work—the way many software companies work or also the example that you mentioned from DeepMind. You really start off with the end-to-end problem and try to understand where you can improve things.

How do you constrain AI in the case of AlphaZero? You have to prescribe and you probably have to make certain decisions based on reinforcement learning—Q-learning—and how you parameterise certain things. It is trial and error. You are still engineering things that have been improved in an iterative fashion and that then at some point delivered the breakthrough. In other words, the increments are there, but they might be hidden somewhere, and what is reported is the end result, which is obviously fascinating. I would advocate more incremental work and sharing the incremental work—how you do this and how you improve this more systematically.

The other part of your question is whether we enter a new area or new paradigm with AI applied to science. There is a change, but the idea of using data is as old as science itself. The empirical method is the idea of using data points. Every field of science has worked with data. Using the image of the toolbox, AI is just another tool in the box that helps scientists to do their work sometimes more efficiently.

The example I gave to the question earlier was of medical interactions in terms of side effects. While the discovery might be groundbreaking, there is still this research base or the way you formulate and test the AI—what the AI is optimising. In the case of AlphaZero, it is two players playing a game against each other and the set-up is done by a human. I have seen very rare cases, except the one example I mentioned about the mathematical intuition, where I would genuinely see this as new ideas that come from AI rather than a more efficient search.

Q202 **Chair:** But to specify a hypothesis and to be able to write it down is necessarily a simplification. It is literally a simplification: you are distilling it into a testable proposition. It may be the case that the combination of interactions is so complex that it defies that simplification, or, certainly, the requirement to simplify it and specify it may limit the power of the multiple connections in the data that can be discerned by machine learning. Is there not a requirement, therefore, to specify hypotheses in advance? Is that not the equivalent of passing the Red Flag Act—that you had to wave a red flag in front of a motor vehicle—in that we are trying to impose current constraints, or constraints that reflect our current worries, on a technology that is transformative?

Professor Vollmer: Probably there are other standpoints. My personal standpoint here is that it is what we define a hypothesis to be. A hypothesis can be very narrow or very broad. It could be saying that the hypothesis is that a diabetes drug might interact with chemotherapy X, and that is a very specific hypothesis. The hypothesis could be that these kinds of AI methods might be able to detect interactions between drugs that have not been known and we feed these existing known interactions in, and this is the kind of area and how we specify the area.

You can formulate higher levels of hypothesis that are maybe not as prescriptive as they might seem. That is the way to go. As I said before, interesting aspects can come while answering this hypothesis. I am not saying that this should limit science and we should not try to do other things. It is just very helpful to be transparent, and it would help the reproducibility crisis a lot if it were clear what is being done from the start, and not being up front that you have tried a million things.

Q203 **Chair:** But what is wrong with trying a million things and finding the one thing that came out? It is virtually impossible for humans to do it, but machines now can. If that one thing saves millions of lives, is that so unrespectable?

Professor Vollmer: I may have been misunderstood. I am not saying it is unrespectable. We just have to make sure that one is very up front about how it is done and which way the million things are checked, and not that it just happens to be that 2 million things are checked, 1 million do not work, the other million do work, and we only report the 1 million. The way to prevent this from happening is to be very up front about what is being checked and what is being looked for, and also being very transparent about the hit rate at the end. Maybe there is one hit-case and you could have found it otherwise, or maybe 99% of it is garbage and it is not much more useful than the status quo, and you have to be very up front about it.

Q204 **Chair:** There are two different aspects, are there not? One is whether it is literally reproducible: that the one-in-a-million hit rate that you have established is reliably so and you can replicate it and find that again. That clearly is very important for safety and confidence in what you found. We have heard in evidence that there is a certain disapproval of testing hypotheses that have not really been thought out, which in this case would have been generated almost at random. That is a different thing, is it not? You want to be able to replicate it, but if you had no idea that something in advance could have had a connection but you have discovered one, that is worth prizing, is it not?

Professor Vollmer: Yes, there is no disagreement.

Q205 **Chair:** Your solution, the way of squaring the circle so that researchers can be open, is to lodge a hypothesis at a very high level that AI techniques and large datasets will be used to make connections for which there may not be any theoretical understanding in advance, and to be up front that you are doing that rather than perhaps ascribing some personal piece of brilliance to the result that then comes out that you retrofit as being your own idea in the first place. Is that what you are recommending?

Professor Vollmer: I am not quite sure I completely understood. Could you rephrase it? Sorry about that. It is an important point.

Q206 **Chair:** Your suggestion was to specify a broad hypothesis and to be clear about what you are doing so that you say that you may not have any

particular theory as to what two variables are going to interact with each other, but you are going to take a big dataset, you are going to apply AI techniques, and that is your research method. You are specifying at a very high level rather than finding a result and then ex post thinking up a hypothesis that would be consistent with that result.

Professor Vollmer: Yes. To elaborate, this has to be specified a bit more than we discussed. If you say AI methods, part of the registered report would be to detail the steps of developing the AI methods and how the different parts of the AI methods interact.

I think there is no doubt that AI will allow us to answer questions that we could not without it. One specific example that comes to mind is the excess mortality in Yemen using satellite data. You could not have done it with human working and classical methods. It is important and helps us understand the mechanism of the disease in certain countries. Nobody questions that.

If the registered reports help to take a step back and produce more quality than quantity, that is what is needed. I would not say everything has to be a registered report, but more incentives to go down this route would be a big start. It would not limit going down other routes and having accidental discoveries, describing them and later setting up proper processes to validate them. I do not think there is a black and white. We do not want to say everything should be registered reports. It is just that, currently, we have too few of them, and it would benefit the field a lot if they were mandated and valued more.

Chair: Thank you very much. That is very clear. Professor Vollmer, I am very grateful for your evidence today. You have given some real expertise on a very focused part of our inquiry.